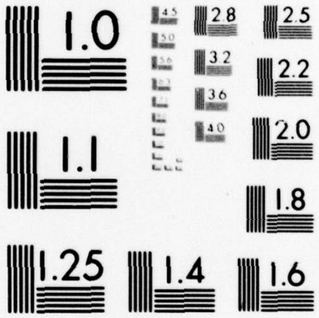
AD-A069 456 CLEMSON UNIV S C DEPT OF MATHEMATICAL SCIENCES F/G 12/1
CHI-SQUARE TESTS FOR THE MULTINOMIAL DISTRIBUTION.(U)
DEC 78 K ALAM N00014-75-C-0451
UNCLASSIFIED N104 NL

| OF |
AD
A069456

END
DATE
FILMED
7-79
DDC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

12 LEVEL II

11 5 Dec 78

12 10p.

6 CHI-SQUARE TESTS FOR THE MULTINOMIAL DISTRIBUTION.

10 KHURSHEED ALAM

DEPARTMENT OF MATHEMATICAL SCIENCES
CLEMSON UNIVERSITY

9 TECHNICAL REPORT 296

14 REPORT N104, TR-296

Research Supported by
The Office of Naval Research
Task NR 042-271 Contract N00014-75-C-0451 15

407 183

CHI-SQUARE TESTS FOR THE MULTINOMIAL DISTRIBUTION

Khursheed Alam

Clemson University

A simple proof of the asymptotic property of Chi-square tests, commonly used in the analysis of categorical data, is given for use as a note for instruction to first-year graduate students.

Key Words: Multinomial Distribution; Chi-square Tests; Contigency Tables.

1. Introduction

The Chi-square tests associated with the multinomial distribution are commonly used in the analysis of categorical data with reference to problems of specification, homogeneity of parallel samples, independence of attributes, etc. The asymptotic property of the tests, that is, the Chi-square distribution of the test statistics in large samples is generally known. However, it has been our observation that many applied statisticians tacitly accept the asymptotic result without satisfying themselves with its proof. It is also true that nearly all the text books in use on elementary and higher statistics either omit the proof or barely sketch it. In this paper we outline a fairly simple proof of the fundamental result, for use as a note for the instruction to first-year graduate students and as a needed theory for the frequent application of Chi-square tests by applied statisticians. The proof is essentially based on the contents of Chapters 5 and 6 of Rao (1966).

Consider a multinomial distribution $M(\underset{\sim}{n},p)$ with K cells, where $p = (p_i,\ldots,p_K)'$ denotes the vector of cell probabilities, $\underset{\sim}{n} = (n_1,\ldots,n_K)'$ denotes the vector of cell frequencies resulting from n independent trials, $\sum_{i=1}^{K} p_i = 1$ and $\sum_{i=1}^{n} n_i = n$. A general problem in judging goodness of fit is to test whether the cell probabilities are specified functions of a fewer number of parameters whose values may be unknown. Let the cell probabilities be given functions $p_1(\theta),\ldots,p_K(\theta)$ of an unknown vector $\theta' = (\theta_1,\ldots,\theta_r)$, where $r < K$. To test the specification, it is a standard method to use the statistic

$$(1.1) \qquad T = \sum_{i=1}^{K} (n_i - np_i(\hat{\theta}))^2 / np_i(\hat{\theta})$$

where $\hat{\theta}$ is a consistent estimator of θ, usually the maximum likelihood estimator. Next, assuming that the specification is true, consider the hypothesis that θ is given by $\theta_i = g_i(\alpha)$, where $g_1,\ldots,g_r$ are given functions, $\alpha' = (\alpha_1,\ldots,\alpha_s)$ and $s < r$. This hypothesis arises in a test of homogeneity of parallel samples and of independence in a contingency table. The statistic

$$(1.2) \qquad T^* = n \sum_{i=1}^{K} (p_i(\hat{\theta}) - p_i(\hat{\alpha}))^2/p_i(\hat{\alpha})$$

is used to test the hypothesis, where $\hat{\alpha}$ denotes an estimate of α and $p_i(\alpha)$ denotes the value of $p_i(\theta)$ as a function of α, under the given hypothesis. It is shown below that T and T* are asymptotically distributed for large n according to the Chi-square distribution under certain conditions.

2. Asymptotic Distribution of T and T*

First, consider the specification that the multinomial cell probabilities are given functions $p_1(\theta),\ldots,p_K(\theta)$ of an unknown vector $\theta = (\theta_1,\ldots,\theta_r)'$, where $r < K$. Let θ^o denote the true value of θ. We make the following assumptions:

(i) The functions $p_i(\theta)$ are continuous in θ, admitting first order partial derivatives which are continuous at θ^o.

(ii) Given $\delta > 0$, there exists $\varepsilon > 0$ such that $\inf_{|\theta-\theta^o|>\delta} N(\theta) > \varepsilon$,

where

$$N(\theta) = \textstyle\sum_{i=1}^K p_i(\theta^o) \log (p_i(\theta^o)/p_i(\theta)).$$

Let

$$M(\theta) = \textstyle\sum_{i=1}^K \frac{n_i}{n} \log (p_i(\theta^o)/p_i(\theta)).$$

Consider the function $N(\theta)$ on the sphere $|\theta-\theta^o| = \delta$. Since $N(\theta)$ is continuous in θ, the infimum of $N(\theta)$ is attained on the sphere. Therefore, in view of (ii), $N(\theta) \geq \varepsilon$ for every point on the sphere. Since $\frac{n_i}{n}$ converges in probability to $p_i(\theta^o)$ as $n \to \infty$, it follows that $M(\theta) > 0$ for all points on the sphere with probability approaching 1 as $n \to \infty$.

The log likelihood function is proportional to $\sum_{i=1}^K n_i \log p_i(\theta)$. In view of (i) and the result given above, we have that for sufficiently large n, the likelihood function has a local maximum inside the open sphere $|\theta-\theta^o|<\delta$ at a point $\hat{\theta}$, say, which is a solution of the likelihood equation

$$\sum_{i=1}^K \frac{n_i}{p_i(\theta)} \frac{\partial p_i(\theta)}{\partial \theta_j} = 0, \quad j = 1,\ldots,r. \tag{2.1}$$

Since δ can be made arbitrarily small, the maximum likelihood estimator $\hat{\theta}$ is a consistent solution of the likelihood equation.

Let $I(\theta) = (M'(\theta))\,(M(\theta))$ denote the information matrix of the multinomial distribution, where

$$M(\theta) = \left(\frac{1}{\sqrt{p_i(\theta)}} \frac{\partial p_i(\theta)}{\partial \theta_j} \right)$$

is a K x r matrix, and let $Z = M'V$, where $M = M(\theta^o)$ and

$$V' = \left(\frac{n_1 - np_1(\theta^o)}{\sqrt{np_1(\theta^o)}}, \ldots, \frac{n_K - np_K(\theta^o)}{\sqrt{np_K(\theta^o)}} \right).$$

By the central limit theorem, the asymptotic distribution of V is multivariate normal $N(\underset{\sim}{0}, I-\phi\phi')$, where $\underset{\sim}{0}$ denotes the null vector, I denotes the identity matrix and $\phi' = (\sqrt{p_1(\theta^o)}, \ldots, \sqrt{p_K(\theta^o)})$. The asymptotic distribution of Z is $N(\underset{\sim}{0}, I)$, where $I = I(\theta^o)$. Note that $M'\phi = \underset{\sim}{0}$.

Substituting $\hat{\theta}$ for θ in (2.1), the jth equation can be written as

$$(2.2) \qquad \sum_{i=1}^{K} \frac{n_i - np_i(\theta^o)}{\sqrt{n}p_i(\hat{\theta})} \frac{\partial p_i(\hat{\theta})}{\partial \hat{\theta}_j} = \sum_{i=1}^{K} \frac{\sqrt{n}(p_i(\hat{\theta}) - p_i(\hat{\theta}^o))}{p_i(\hat{\theta})} \frac{\partial p_i(\hat{\theta})}{\partial \hat{\theta}_j}.$$

In view of (i) we have

$$(2.3) \qquad p_i(\hat{\theta}) - p_i(\theta^o) = \sum_{\ell=1}^{r} (\hat{\theta}_\ell - \theta_\ell^o) \frac{\partial p_i(\theta^o)}{\partial \theta_\ell^o} + \eta|\hat{\theta} - \theta^o|$$

where $\eta \to 0$ as $\hat{\theta} \to \theta^o$. Since $\hat{\theta}$ is a consistent estimator of θ, as shown above, the left side of (2.2) is asymptotically equivalent (converging in probability) to Z_j. Therefore, by the substitution of (2.3) on the right side of (2.2) we have that

$$Z_j \overset{a}{=} \sum_{\ell=1}^{r} \sqrt{n}(\hat{\theta} - \theta_\ell^o)\, I_{\ell j}$$

where $\overset{a}{\sim}$ means "asymptotically equivalent to", and $I_{\ell j}$ denotes the ℓjth element of I. Hence

$$Z \overset{a}{\sim} \sqrt{n}\ I\ (\hat{\theta} - \theta^{o})$$

or

$$I^{-} Z \overset{a}{\sim} \sqrt{n}\ (\hat{\theta} - \theta^{o}) \tag{2.4}$$

where I^{-} denotes a generalized inverse of I, given by $I\ I^{-}\ I = I$, and is equal to I^{-1} if I is non-singular.

Let A be a symmetric matrix with real elements, and let $X \overset{d}{\sim} N\ (\mu,\Sigma)$, where $\overset{d}{\sim}$ means "distributed as". If the covariance matrix Σ is non-singular then it is known that the quadratic form $X'\ A\ X \overset{d}{\sim} \chi^{2}_{\nu,\delta}$ (non-central Chi-square with ν degrees of freedom and non-centrality parameter δ) if and only if $A\ \Sigma$ is idempotent, where ν = Rank A and $\delta = \frac{1}{2}\mu' A\mu$. If Σ is singular, then the given condition is only sufficient and ν = Rank $A\Sigma$ (see e.g. Graybill (1976), Theorem 4.7.1). If $A = \Sigma^{-}$ is a generalized inverse of Σ, given by $\Sigma\ \Sigma^{-}\Sigma = \Sigma$, then $A\ \Sigma$ is idempotent and Rank $A\ \Sigma$ = Rank Σ. Therefore, $X'\ \Sigma^{-}\ X \overset{d}{\sim} \chi^{2}_{\nu,\delta}$, where ν = Rank Σ and $\delta = \frac{1}{2}\ \mu'\ \Sigma^{-}\ \mu$.

Let $W = (W_{1},\ldots,W_{K})'$ where

$$W_{i} = \sqrt{n}\ (p_{i}(\hat{\theta}) - p_{i}\ (\theta^{o}))/\sqrt{p_{i}(\theta^{o})}\quad ,\ i = 1,\ldots,K.$$

From (2.3) we have that

$$\begin{aligned} W &\overset{a}{\sim} \sqrt{n}\ M\ (\hat{\theta} - \theta^{o}) && (2.5)\\ &\overset{a}{\sim} M\ I^{-}\ Z \qquad \text{by (2.4)}\\ &= M\ I^{-}\ M'\ V. \end{aligned}$$

Note that $M I^{-} M'$ is idempotent. From (1.1) and (2.5) we have

$$T \overset{a}{\approx} (V - W)' (V - W)$$

$$\overset{a}{\approx} V' (I - M I^{-} M')V.$$

Now, V is asymptotically distributed as $N(\underset{\sim}{0}, I - \phi\phi')$, $(I - MI^{-}M')(I-\phi\phi')$ $= I - MI^{-}M'-\phi\phi'$ is idemptotent and

$$\text{Rank } (I-MI^{-}M'-\phi\phi\ ) = \text{Trace } (I-MI^{-}M'-\phi\phi')$$

$$= K - 1 - \text{Rank } I = \beta\text{, say.}$$

Therefore, T is asymptotically distributed as χ^2_β. If I is of full rank then $\beta = K - r - 1$.

Next, consider the hypothesis that θ is given by $\theta_i = g_i(\alpha)$, $i = 1,\dots,r$, as described in the previous section. Let $\nabla(\alpha) = (\partial\theta_i/\partial\alpha_j)$ denote the r x s matrix of the derivatives, and let $I^*(\alpha)$ denote the information matrix under the given hypothesis. Then

(2.6) $$I^*(\alpha) = (\nabla(\alpha))' I(\theta)\nabla(\alpha)$$

with $I(\theta)$ being expressed as a function of α. Let α^o denote the true value of α. Similarly, as in the preceding we have that

$$T^* \overset{a}{\approx} V'(MI^{-}M'-M\nabla(\nabla'I\nabla)^{-}\nabla'M')V$$

where $\nabla = \nabla(\alpha^o)$. The matrix $(MI^{-}M'-M\nabla(\nabla'I\nabla)^{-}\nabla'M'((I-\phi\phi')=(MI^{-}M'-M\nabla(\nabla'I\nabla)^{-}\nabla'M')$ is idempotent and

$$\text{Rank } (MI^{-}M'-M\nabla(\nabla'I\nabla)^{-}\nabla'M') = \text{Rank } I\text{-Rank } (\nabla'I\nabla)$$

$$= \gamma\text{, say.}$$

Therefore, T* is asymptotically distributed as χ_{γ}^{2} . If I and ∇ are of full rank then γ = r-s.

We have shown that T and T* are asymptotically distributed according to the Chi-square distribution under the identifiability condition (i) and the continuity assumption (ii).

Remark: For the goodness of fit test where the cell probabilities are completely specified we have β = K-1. In this case $T \overset{a}{\underset{\sim}{=}} V'V \overset{d}{\sim} \chi_{K-1}^{2}$, asymptotically. For testing homogeneity of r parallel samples or independence of attributes in r x K contigency tables, we have γ = (r-1) (K-1).

References.

[1] Graybill, R. A. (1976). The Theory and Application of the Linear Model. Wadsworth Publishing Co., Belmont, California.

[2] Rao, C. R. (1966). Linear Statistical Inference and its Applications. Wiley Publications in Statistics.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
N104

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Chi-square tests for the multinomial distribution

5. TYPE OF REPORT & PERIOD COVERED
Technical Report

6. PERFORMING ORG. REPORT NUMBER
296

7. AUTHOR(s)

8. CONTRACT OR GRANT NUMBER(s)
N00014-75-C-0451

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Clemson University
Dept. of Mathematical Sciences
Clemson, South Carolina 29631

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
NR 042-271

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
Code 436
Arlington, Va. 22217

12. REPORT DATE
12-5-1978

13. NUMBER OF PAGES
7

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Multinomial distributions; Chi-square test; Contingency table.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper gives a proof of the asymptotic property of the Chi-square tests associated with the multinomial distribution, generally used in the analysis of categorical data, such as contingency tables.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)